结合改进的 CHI 统计方法的 TF-IDF 算法优化 *

马　莹，赵　辉†，李万龙，庞海龙，崔　岩

(长春工业大学 计算机科学与工程学院，长春 130012)

摘　要：特征项的选择和特征权值的计算是文本分类过程中两个至关重要的环节，对文本分类的结果起关键性作用。为了克服传统的 CHI 统计方法存在特征项出现频率与类别负相关的情况和某一个特征项存在于某一个文本中的概率问题，针对传统的 CHI 统计方法引入了负相关判定、频度等重要因素进行了改进，并结合语义相似度的计算方法对 TF-IDF 算法进行了优化，在 WEKA 软件上采用了 KNN（K-nearest neighbor）分类器和支持向量机（SVM）分类器分别对微博情感语料进行分类，该实验结果表明，新方法在文本分类的准确性上有明显的提高。

关键词：文本分类；CHI 统计；TF-IDF 算法；特征选择

中图分类号：TP301.6　　doi: 10.3969/j.issn.1001-3695.2018.01.0136

Optimization of TF-IDF algorithm combined with improved CHI statistical method

Ma Ying, Zhao Hui†, Li Wanlong, Pang Hailong, Cui Yan

(*College of Computer Science & Engineering*, *Changchun University of Technology*, *Changchun* 130012, *China*)

Abstract: The selection of feature items and the calculation of feature weights are two crucial links in the process of text classification and play a key role in the results of text classification. In order to overcome the traditional CHI statistical method, there is a negative correlation between the frequency of feature items and the category, and a probability problem that a feature item exists in a text, The traditional CHI statistical method is improved by introducing some important factors such as negative correlation judgment and frequency, and the TF-IDF algorithm is optimized by combining the calculation method of semantic similarity. The K-nearest neighbor (KNN) classifier and support vector machine (SVM) classifier are respectively used in WEKA software to classify the Weibo emotional corpus The experimental results show that the new method has obvious improvement on the accuracy of text classification.

Key words: text categorization; CHI statistics; TF-IDF algorithm; feature selection

0　引言

随着互联网的快速进步，电子信息量越来越膨胀，那么如何将大量的信息进行有规律的、有效的、有组织的管理呢[1,2]？文本分类作为处理和归纳大量数据的关键技术，可以在很大程度上解决信息没有条理的现象，使用户可以从大量的信息中快速得获取有价值的信息，因此在舆情控制、信息安全、协同过滤、产品推荐等方面具有广泛应用[3]。

在文本分类过程中有两个主要的因素影响着分类的最终结果：一是特征项的选择，二是特征项权重的计算方法。特征选择是从大量的词语中选择出一部分有价值的特征项进行分类，使分类的结果达到最优化[4]。特征项的权重计算方法是用于文本的数据统计，给特征项赋权，特征权重是用来权衡某一个特征项在某一个文本中的重要程度[5]。

目前，文本分类研究方面：郭正斌通过利用权值和维度的调整对文本向量空间模型进行优化，提出了一种新的面向文本分类的优化方法，可以达到优化向量空间目的[6]。周庆平将改进的 X^2 统计方法与聚类相结合，最后通过改进的 KNN 进行分类，可以提高分类效果[7]。徐明针对于微博特征选择的问题作了相关的研究与说明，提出了一种新的卡方统计的算法，在 KNN 和 SVM 分类下进行测试，得出了在微博信息分类上准确率有一定的提高[8]。

文本分类过程中最常用的特征选择方法之一是 CHI 统计方法，传统的 CHI 统计方法没有考虑到特征项出现频率与类别负相关的问题，也没有顾虑到某一个特征项存在于某一个类别文本的情况，为解决以上问题，本文通过判断正负的方式去除了特征项出现频率与类别负相关的情况，并引入了频度等重要因素对传统的 CHI 统计方法进行了相应的改进，在此基础上引入

收稿日期：2018-01-27；**修回日期：**2018-04-10　　**基金项目：**国家自然科学基金资助项目(61472049)；吉林省教育厅“十二五”科学技术研究项目(2014132)

作者简介：马莹（1993-），女，吉林长春人，硕士研究生，主要研究方向为自然语言处理、智能计算；赵辉（1972-），女（通信作者），教授，博士，主要研究方向为智能计算、搜索引擎（412600729@qq.com）；李万龙（1963-），男，教授，博士，主要研究方向为软件工程与智能系统；庞海龙（1987-），男，硕士研究生，主要研究方向为推荐系统、智能计算；崔岩（1993-），男，硕士研究生，主要研究方向为推荐系统、智能计算.

了特征提取因子 ε，将语义相似度算法与传统的 TF-IDF 算法进行了结合优化，从而提高某一个特征项在类别文本中的重要程度，达到降低维度的效果，最终提高了文本分类的准确性。

1　CHI 统计方法

CHI 统计方法是用来测量特征项 w 与类别 c 之间的相关性，其关联列表如表 1 所示。

表 1　特征项与类别关系表

类别 / 特征项	K	$\overline{K}$	总计
w	A	B	A+B
$\overline{w}$	C	D	C+D
总计	A+C	B+D	A+B+C+D

其中，属于类别 K 且存在特征项 w 的文本数量，命名为 *A*；不属于类别 K 但存在特征项 w 的文本数量，命名为 *B*；属于类别 K 但不包含特征项 w 的文本数量，命名为 *C*；不属于类别 K 也不包含特征项 w 的文本数量，命名为 *D*。则 CHI 值的计算公式如下：

$$X^2(w,c)=\frac{N(AD-BC)^2}{(A+C)(A+B)(B+D)(C+D)} \quad (1)$$

且有 N=A+B+C+D。

当 CHI 的值等于 0 时，表示特征项 w 与类别 c 之间没有任何关系；当 CHI 的值越大时，表示特征项 w 与类别 c 的关系性越强。

现有研究发现，传统 CHI 统计有两处不足：

a）传统的 CHI 统计方法存在着特征项出现频率与类别负相关的问题，还存在着倾向于选取在文本中出现比例相对较少的特征项，这样的特征项中的绝大部分与类别没有较强的联系，甚至没有任何关系，只有个别的特征词与类别存在较强的联系。

b）传统的 CHI 统计方法没有考虑到某一个特征项存在于某一个文本中的概率，而只是考虑了存在于全部文本中的概率，如果某一个特征项在某一类别的多数文本中集中存在，而在此类别的少数文本中很少存在，则此时 CHI 的值可能会高，相反，CHI 值可能会低。

2　改进的 CHI 统计方法

特征项与文本类别存在两种关系情况：

a)当 AD-BC 的值大于 0 为正数时，说明特征项出现频率与类别为正相关，即特征项存在的文本属于某一个类别的机率越大，正数越大，那么它的平方越大，则 CHI 的值就越大，因此可以作为特征选择的特征项。

b)当 AD-BC 的值小于 0 为负数时，说明特征项出现频率与类别为负相关，即特征项出现的文本属于某一个类别的机率越小，负数越小，那么它的平方越大，则 CHI 的值就越大，因此不可以作为特征选择的特征项。

从传统的 CHI 统计方法的计算公式可以看出，如果特征项出现频率与类别负相关的问题没有解决，那么这种负相关的情况最后会影响 CHI 的值，所以特征选择的结果会受到影响，进而影响到文本分类的精确率，因此本文通过去除特征项出现频率与类别负相关的方式来解决这一问题，改进后的公式为

$$X^2(w,c)=\begin{cases}\dfrac{N(AD-BC)}{(A+C)(A+B)(B+D)(C+D)} & AD-BC>0\\ 0 & AD-BC<0\end{cases} \quad (2)$$

由于 CHI 统计方法没有考虑某一个特征项存在于某一个文本中的次数，而只是考虑了存在于全部文本中的次数，文献[9]中指出特征选择的关键在于特征项集中均匀的分布在某一类别文本中，所以本文将频度 α，集中度 β，分散度 γ 引进传统的 CHI 统计公式中。假设训练集类别为 C_j 的文本有 $d_{j1},...,d_{jk},...d_{jm}$，$tf_{jk}$ 表示特征项 w 在文本 d_{jk}（$1\leqslant k\leqslant m$）中存在的频率，m 表示某个类别中的文本总数，df_j 表示类 C_j 中含有特征项 w 的本文数量，n 表示文本类别总数量。

a)频度，指某一类别文本中出现某一个特征项的次数占此类别的文本总数量的比重。则特征项 w 在类别 Cj 中存在的频度 α 表示如下：

$$\alpha=\sqrt{\sum_{k=1}^{m}(tf_{jk})^2} \quad (3)$$

b)集中度，指某一个类别文本中存在的某一个特征项的文本数量占含此特征项的文本数量的比重。则特征项 w 存在于类别 C_j 中的集中度 β 表示如下：

$$\beta=\frac{(n\cdot df_j-\sum_{j=1}^{n}df_j)^2}{n\cdot\sum_{j=1}^{n}df_j} \quad (4)$$

c)分散度，某一个类别中含某一个特征项的文本数量占该类别量中总文本数的比重。则分散度 γ 表示如下：

$$\gamma=\frac{df_j}{m} \quad (5)$$

由以上的定义可知，某一个特征项 w，聚集的存在于某一个类别的大部分文本中的次数越多，说明其频度越高，集中度强，分散度大，这样的特征项对文本分类结果有很大的作用，于是在式（2）的基础上引入了频度、集中度、分散度，得到如下公式：

$$X^2(w,c)=\begin{cases}\dfrac{N(AD-BC)}{(A+C)(A+B)(B+D)(C+D)}\times\alpha\times\beta\times\gamma & AD-BC>0\\ 0 & AD-BC<0\end{cases} \quad (6)$$

3　传统的 TF-IDF 算法及其改进

3.1　传统的 TF-IDF 算法

TF-IDF 通常用于衡量一个字或者一个词语在语料集中的重要程度。TF-IDF 算法是由 Jones[10]首次提出的，TF-IDF 实际上是 TF 与 IDF 的乘积。

$$TF=\frac{m}{M} \quad (7)$$

其中：m 代表存在于文本 i 中的特征项的次数，M 代表文本 i 中

的总词语数量。

$$\mathrm{IDF}=\log\left(\frac{\mathrm{N}}{\mathrm{n}}+0.01\right) \tag{8}$$

其中:N为总文本数量，n为含有某个特征项的文本总数量。

则 TF-IDF 的特征提取函数为

$$\mathrm{F}(w)=TF(w)\bullet IDF(w) \tag{9}$$

归一化后的传统 TF-IDF 公式如下：

$$W_{ij}=\frac{tf_{ij}\times\log\left(N/n_{ij}+0.01\right)}{\sqrt{\sum_{j=1}^{M}\left[tf_{ij}\times\log\left(N/n_{ij}+0.01\right)\right]^2}} \tag{10}$$

其中:W_{ij}代表特征项的权重，tf_{ij}代表某一个特征项存在于某一个文本中的频率，$log(N/n_{ij}+0.01)$为逆文本频率，N 代表类别文本的总数量，n_{ij}代表存在某一个特征项的文本数量。

3.2　优化的 TF-IDF 算法

采取了传统的归一化 T 特征项与类别变化方向相反 F-IDF 算法来给特征项赋权值时，只考虑了特征项在分本中的分布情况，并没有考虑词语的近义词存在于文本中情况，忽略了词语之间的相似性，如果采取该算法给特征项赋权值就忽略了文本中的这一特点。文献[11]中提出了词语相似度的计算方法，通过对知识语言分析可知,知识语言的数据结构可以用集合与义原、特征结构来表达，语义的相似度计算是采取“知网”中的计算相似度的算法来确定，γ 的系统设定值为 0.8。该算法对词语相似度的准确性有一定的提高。为了解决特征词在文本中会出现近义词的情况，本文将语义相似度的计算方法应用到传统的 TF-IDF 算法中，从而使特征项在文本中的权重变大，让此特征项更具有代表性意义。为此本文提出了特征提取因子 ε， ε 代表出现在某一个文本中的某一个特征项的数量与此特征项相似度大于 γ 的特征项的数量之和与所有特征项的数量的比例， ε 的结果直接影响到特征项在文本中的重要程度。其定义公式如下：

$$\varepsilon=\frac{a+b}{c} \tag{11}$$

其中:a 代表存在于文本 i 中的特征项 t_{ij}的数量，b 代表与特征项 t_{ij}相似度大于 γ 的特征项的数量，c 代表所有特征项的数量。为了提高特征项权值的准确度，本文在结合语义相似度算法的基础下引进了特征提取因子 ε 对传统的归一化 TF-IDF 算法进行了优化，实现了形式与语义的结合。其定义公式如下：

$$\mathrm{W_{ij}}=\frac{\sqrt{tf_{ij}\times\varepsilon}\times\log\left(N/n_{ij}+0.01\right)}{\sqrt{\sum_{j=1}^{M}\left[\sqrt{tf_{ij}\times\varepsilon}\times\log\left(N/n_{ij}+0.01\right)\right]^2}} \tag{12}$$

其中:W_{ij} 代表特征项的权重，N 代表类别文本的总数量，n_{ij} 代表某一个特征项的文本数量与此特征项的相似度大于 γ 的特征项的文本数量之和的平均值。

4　实验结果与分析

4.1　实验数据及环境

本文采用微博文本作为实验数据，与传统的网络文本进行对比，微博文本较短，对字数有严格的要求约束，规定以 140 个字符为限，并且具有偶然性，实时性，不可控性的特点[12]。实验数据采用了 4000 条的新浪微博语料进行数据分析，电脑系统是 Window 7，软件是 python 编程技术，采用 Weka3.6 数据挖掘开源工具进行实验结果对比。在实验中分别使用了 KNN 和 SVM 分类器进行数据测试分析。

4.2　评价指标

假定在类别的分类结果中，X 代表某一个特征项被正确分类到某一个类别的文本数量，Y 代表某一个特征项被错误分类到某一个类别的文本数量，Z 代表某一个特征项被遗漏分类的文本数量。具体的公式表示如下：

$$\text{召回率}\mathrm{R}=\frac{\mathrm{X}}{\mathrm{X}+\mathrm{Z}}\times 100\% \tag{13}$$

$$\text{查准率}\mathrm{P}=\frac{\mathrm{X}}{\mathrm{X}+\mathrm{Y}}\times 100\% \tag{14}$$

$$\mathrm{F}\text{值}=\frac{2\times P\times R}{P+R}\times 100\% \tag{15}$$

4.3　实验及结果

实验 1　将微博的情感分为两方面，一是正向情感，二是负向情感，比较了在相同维度的特征下，对结合改进的 CHI 统计方法的 TF-IDF 算法、传统的 CHI 统计方法和改进的 CHI 统计方法在 KNN 分类器效果下的进行了实验对比。实验结果如表 2 所示。

表 2　500 维度 KNN 分类器下三种方法的对比

语料	传统的 CHI			改进的 CHI			结合改进的 CHI 和优化的 TF-IDF		
指标	R	P	F	R	P	F	R	P	F
正向情感	89.7	85.4	87.5	89.5	87.0	88.2	90.0	88.5	89.2
负向情感	86.1	90.2	88.1	87.4	89.8	88.6	88.0	89.7	88.8
平均	87.9	87.8	87.8	88.4	88.4	88.4	89.0	89.1	89.0

从表中可以看出,在相同维度的特征下采用 KNN 分类器，通过对微博的正向情感和负向情感在召回率 R、查准率 P、F 值三个指标上进行了比较，得出结合改进的 CHI 统计方法的 TF-IDF 算法比传统的 CHI 统计方法和改进的 CHI 统计方法在三个指标上都有一定的提升,结合改进的 CHI 统计方法的 TF-IDF 算法的平均查准率比传统的 CHI 统计方法的平均查准率高出 1.3 个百分点，说明结合改进的 CHI 统计方法的 TF-IDF 算法提高了微博情感分类的准确率。

实验 2　比较了在不同维度下，对结合改进的 CHI 统计方法的 IF-IDF 算法、传统的 CHI 统计方法和改进的 CHI 统计在 KNN 分类效果下的进行了实验，对以上三者在查准率上进行了对比。实验结果如表 3 所示。

表 3　不同维度下 KNN 分类器三种方法的对比

维度	传统的 CHI	改进的 CHI	结合改进的 CHI 和优化的 TF-IDF
400	87.7	88.4	89.0
600	88.0	89.0	89.3
800	89.0	90.0	90.8
1000	89.5	90.2	91.3

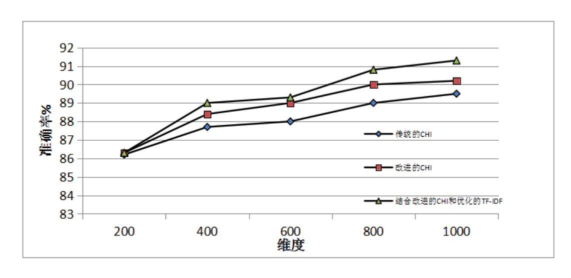


图 1　在不同维度下 KNN 分类器的准确率

从图 1 中可以看出：无论是在哪一个维度特征下采用 KNN 分类器，结合改进的 CHI 统计方法的 TF-IDF 算法在查准率 P 上比传统的 CHI 统计方法和改进的 CHI 统计方法的分类效果更为可观，特别是在 200 和 600 维度的特征下此算法的准确率有明显的提高。

实验 3　比较了在 KNN，SVM 分类器下，维度分别在 200、400、600、800、1000 维度下的传统的 CHI 统计方法、改进的 CHI 统计方法和结合改进的 CHI 统计方法的 TF_IDF 算法三者之间的性能，实验结果如表 4 所示：

表 4 在不同维度 SVM 分类器下三种方法的对比

维度	传统的 CHI	改进的 CHI	结合改进的 CHI 的 TF-IDF 优化
200	87.5	88.0	88.5
400	87.9	88.2	90.5
600	88.5	89.5	90.8
800	89.4	90.6	91.9
1000	90.3	91.2	92.5

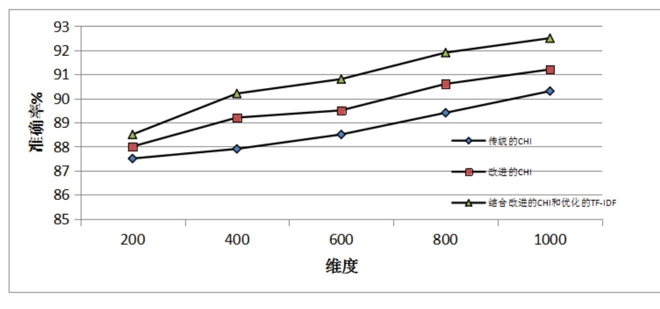


图 2　不同维度下 SVM 分类器的准确率

将图 2 与图 1 的实验结果进行对比可以看出，在相同的维度下，结合改进的 CHI 统计方法的 TF-IDF 算法采用 SVM 分类器的分类效果比 KNN 分类器的分类效果好，此结果与其他有关微博研究的结果基本一致。

4.4　实验结果分析

本文主要研究结合改进的 CHI 统计方法的 TF-IDF 算法的优化，通过对传统的 CHI 统计方法进行了相应的改进，并与结合语义相似度的传统的 TF-IDF 算法进行了结合优化。实验的结果显示：在表 1 中可以看出在 500 维度下采用 KNN 分类器，改进 CHI 统计方法比传统的 CHI 统计方法的平均查准率高出 0.6 个百分点，而本文提出的结合改进的 CHI 统计方法的 TF-IDF 的算法比传统的 CHI 统计方法的平均准确率高出 1.3 个百分点；在表 2 中可以看出在 400 维度下结合改进的 CHI 统计方法的 IF-IDF 算法的优化的准确率上升幅度比较小，产生这种原因的可能是因为某一个词语的近义词的影响，在 600 维度下准确率又得到了提高；将表 2 与 3 的实验结果数据进行对比可以看出：在相同的维度特征下，采用结合改进的 CHI 统计方法的 TF-IDF 算法，与 KNN 分类器相比，SVM 分类器的分类效果更为可观，说明 SVM 分类器更适合于文本提出的新方法，最终达到了提高微博情感分类准确率的目的。

5　结束语

通过对文本分类技术的研究与试验，本文提出了结合改进的 CHI 统计方法的 TF-IDF 算法优化。首先针对于传统的 CHI 统计方法进行了相应的改进，弥补了传统的 CHI 统计方法特征项出现频率与类别负相关的问题和某一个特征项存在于某一个文本中的概率情况，然后与结合语义相似度算法的 TF-IDF 算法结合优化，从而提高某个特征项在文本中的重要程度，并达到了降维的效果，最终提高了文本分类的准确性。实验结果分析表明，选择结合改进的 CHI 统计方法的 TF-IDF 算法在 SVM 分类器下进行分类，取得了良好的分类效果，提高了分类的准确率。

参考文献：

[1] 高岩. 微博情感分析的相关技术研究 [D]. 北京：华北电力大学, 2014. (Gao Yan. Research on related technologies of Weibo sentiment analysis [D]. Beijing: North China Electric Power University, 2014.)

[2] 万源. 基于语义统计分析的网络舆情挖掘技术研究 [D]. 武汉：武汉理工大学, 2012. (Wan Y. Research on Internet Public Opinion Mining Technology Based on Semantic Statistics Analysis [D]. Wuhan University of Technology, 2012.)

[3] 徐燕, 李锦涛, 王斌, 等. 基于区分类别能力的高性能特征选择方法 [J]. 软件学报, 2008, (1): 82-89. (Xu Yan, Li Jintao, Wang Bin *et al.* A high performance feature selection method based on differentiating category capability [J]. Journal of Software, 2008, (1): 82-89.)

[4] 游凤芹, 钟芳, 周展. 中文多类别情感分类模型中特征选择方法 [J]. 计算机应用, 2016, 36 (S2): 242-246. (You Fengqin, Zhong Fang, Zhou Zhan. Feature selection method in Chinese multi-category sentiment classification model [J]. Computer Applications, 2016, 36 (S2): 242-246.)

[5] 王景中, 邱铜相. 基于 TF-IDF 改进算法的聚焦主题网络爬虫 [J]. 计算机应用, 2015, 35 (10): 2901-2904, 2919. (Wang Jingzhong, Qiu Cuxiang. Focused topic web crawler based on improved TF-IDF algorithm [J]. Computer Applications, 2015, 35 (10): 2901-2904, 2919.)

[6] 郭正斌, 张仰森, 蒋玉茹. 一种面向文本分类的特征向量优化方法 [J]. 计算机应用研究, 2017, 34 (8): 2299-2302, 2348. (Guo Zhengbin, Zhang Yangsen, Jiang Yuru. A feature vector optimization method for text classification [J]. Application Research of Computers, 2017, 34 (8): 2299-

2302, 2348.)

[7] 周庆平，谭长庚，王宏君，等. 基于聚类改进的 KNN 文本分类算法 [J]. 计算机应用研究, 2016, 33 (11): 3374-3377, 3382. (Zhou Qingping, Tan Changgeng, Wang Hongjun, *et al.* Improved KNN text classification algorithm based on clustering [J]. Application Research of Computers, 2016, 33 (11): 3374-3377, 3382.)

[8] 徐明，高翔，许志刚，等. 基于改进卡方统计的微博特征提取方法 [J]. 计算机工程与应用, 2014, 50 (19): 113-117, 142. (Xu Ming, Gao Xiang, Xu Zhigang, *et al.* Microblog feature extraction method based on improved chi-square statistics [J]. Computer Engineering and Applications, 2014, 50 (19): 113-117, 142.)

[9] 熊忠阳，张鹏招，张玉芳. 基于χ2 统计的文本分类特征选择方法的研究 [J]. 计算机应用, 2008, 28 (2): 513-514, 518. (Xiong Zhongyang, Zhang Pengzhao, Zhang Yufang. Study of text categorization feature selection method based on χ2 statistics [J]. Computer Applications, 2008, 28 (2): 513-514, 518.)

[10] Jones K S. A statistical interpretation of term specificity and its application in retrieval [J]. Journal of Documentation, 1972, 28 (1): 11-21.

[11] 任姚鹏，陈立潮，张英俊，等. 结合语义的特征权重计算方法研究 [J]. 计算机工程与设计, 2010, 31 (10): 2381-2383, 238 (Ren Yaopeng, Chen Lichao, Zhang Jianjun, *et al.* Research on the calculation methods of feature weights combined with semantics [J]. Computer Engineering and Design, 2010, 31 (10): 2381-2383, 2387.)

[12] 张剑峰，夏云庆，姚建民. 微博文本处理研究综述 [J]. 中文信息学报, 2012, 26 (4): 21-27, 42. (Zhang Jianfeng, Xia Yunqing, Yao Jianmin. Summary of research on microblogging text processing [J]. Journal of Chinese Information Processing, 2012, 26 (4): 21-27, 42.)